B.11. De invisibilidades y cegueras. Los fracasos en la recuperación de información publicada en libros

Por **Carlos B. Amat**

Amat, Carlos B. "De invisibilidades y cegueras. Los fracasos en la recuperación de información publicada en libros". En: *Anuario ThinkEPI*, 2008, pp. 71-75.

Resumen: *La dificultad de recopilar información procedente de registros de bases de datos ha originado la existencia de internet invisible. En los sistemas catalográficos y las bases de datos bibliográficas, la falta de representación del contenido de los libros provoca la ceguera de los usuarios. Los buscadores están desarrollando procedimientos para recopilar y hacer accesibles los contenidos hasta ahora ocultos. Es necesario que los sistemas de registro (Isbn) y los sistemas bibliotecarios tomen iniciativas para desvelar el contenido de los libros, sobre todo los especializados, y hacer accesibles sus componentes a los lectores. Se proponen modificaciones en los sistemas de registro y la elaboración de pasarelas que, a modo de protocolos similares a Z39.50, favorezcan la captura de los capítulos de libros, los sumarios o las sinopsis de las obras de creación para su incorporación a las descripciones bibliográficas de los catálogos automatizados.*

Palabras clave: *Capítulos de libros, Catalogación, Bibliotecas especializadas.*

Title: On invisibilities and blindness: Failures in retrieval of information from books

Abstract: *The difficulty of gathering information from databases not indexed by search engines is the origin of what is called the deep or invisible Internet. In bibliographic and cataloguing systems, lack of description of book contents leaves users with a kind of blindness. Search technologies are being developed that try to index and provide access to this invisible content. Bibliographic and cataloguing systems, such as ISBN, must take measures to gather and process book contents, particularly specialized texts, and make them accessible to readers. It is proposed that these systems be modified to provide gateways, using protocols similar to Z39.50, that support the capture of book chapters, summaries or synopses of literary works so that the information can be incorporated into the corresponding areas of bibliographic records in online catalogues.*

Keywords: *Book chapters, Cataloguing, Special libraries.*

1. "...from nothing to a state of extreme poverty".

LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN en internet, que por comodidad llamamos buscadores, han demostrado sobradamente su ineficacia. Cifras de exhaustividad que no se elevan más allá del 14% a costa de una precisión cuyo mejor resultado no sobrepasa el 30%, o ritmos de actualización tan portentosos que casi el 13% de los resultados conducen a la estrella de la familia de errores 400 (el "No encontrado"), son sólo algunos datos[1]**.**

Estos resultados se obtuvieron en el análisis de los buscadores de tipo sintáctico. Los sistemas basados en el análisis contextual (*Google* el más popular) sin duda presentan mejor rendimiento. Pero todos sin excepción han topado con una serie de limitaciones importantes en la recopilación automática de la información de sedes y páginas y, de paso, han generado una curiosa paradoja.

Seis años después de que *Yahoo* inaugurase los buscadores de información en el espacio web, uno de tantos informes de tipo aguafiestas puso el dedo en la llaga de la limitación antes aludida: existen documentos e información que los buscadores no pueden recopilar. Se trata de información contenida nativamente en registros de bases de datos que sólo es visible a través de la interacción entre los usuarios y los formularios de búsqueda que permiten su selección. Cuantita-

Los buscadores no pueden recopilar información contenida nativamente en registros de bases de datos que sólo es visible a través de la interacción entre los usuarios y los formularios de búsqueda que permiten su selección

tivamente, esta información no es grande, tampoco masiva… es ingente, una auténtica barbaridad: unos 7.500 terabytes, entre 400 y 550 veces más abundante que la información y los documentos accesibles[2].

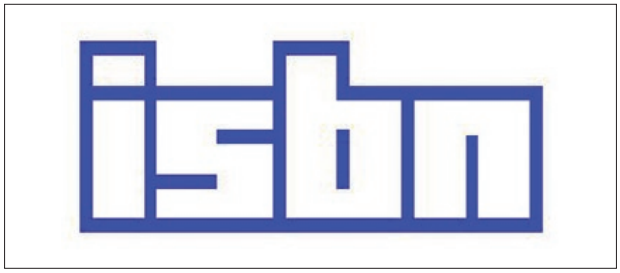
De entre quienes se precipitaron a bautizar el fenómeno, hubo quien abundó en metáforas náuticas, habló de profundidades (*deep Web*) y acompañó la expresión con la inevitable figura del arrastrero cortito de redes. Otros adoptaron terminología y actitud académicas: "Se denomina 'internet invisible' o *Infranet* al conjunto de recursos accesibles únicamente a través de algún tipo de pasarela o formulario web y que, por tanto, no pueden ser indizados de forma estructural por los robots de los buscadores". Y hasta ensayaron una taxonomía cuya primera clase interesa mencionar: "…Catálogos de bibliotecas y bases de datos bibliográficas. Los *opac* y las bases de datos de registros bibliográficos accesibles a través de pasarelas web resultan imposibles de interrogar exhaustivamente por los motores convencionales[3]".

¿No resulta curioso que los registros de bases de datos, incluyendo los registros bibliográficos, permanezcan invisibles a los sistemas aludidos? Esos registros, elaboradas representaciones de documentos que siguen esquemas muy definidos, están ideados y organizados precisamente para su recuperación. Los documentos codificados en mero html, redundantes, inestables, irregulares, de contenidos heterogéneos y no siempre contrastados, tan dinámicos que se les achacaría el baile de San Vito, ésos, sobre todo ésos, son los realmente accesibles a través de los buscadores.

Si algo no se ve, una de dos: o ese algo es invisible o quien intenta percibirlo es ciego. La invisibilidad de la información es un concepto nacido en el marco de la recuperación de información en internet. Pero en lo que se refiere a la recuperación a través de bases de datos de registros estructurados, hay que referirse a las cegueras. Veamos (es un decir) lo que pasa con ellas.

2. Garrapata y John Silver en el país de los ciegos

Supongamos que un auténtico enloquecido por las historias de bucaneros, corsarios, filibusteros y contrabandistas decide dedicar su ocio a las historias de piratas. La consulta del *Isbn*, un mero registro administrativo[4], podría ofrecerle algo más de 300 libros donde elegir. En una base de datos catalográfica[5] la cosecha sería mayor: unos 4.600 títulos. Si nuestro potencial lector es loco, pero un loco bajito, seguro que se siente atraído por la simpática serie de Garrapata. Al fin y al cabo su primera entrega ("El pirata Garrapata", *Isbn* 84-348-1001-8) ha acumulado 36 ediciones y muchos miles de lectores de nueve años. También podría adentrarse en los románticos de la mano de **Walter Scott** ("El pirata", *Isbn* 84-495-0117-2) pero es más dudoso que los recursos mencionados dirijan al usuario hacia obras maestras como "La Isla del Tesoro" o "Jim Botón y Lucas el Maquinista". El intrincado código 087.5 y su traducción (Publicaciones infantiles en general. Libros infantiles y juveniles) no han de resultar de gran ayuda. Ni siquiera permitirán alcanzar la última entrega del Capitán Alatriste (821.134.2-3, se siente). Sólo se puede ver lo que lleva la palabra "pirata" en el título.



La literatura de creación parece interesar sobre todo a las sociedades de gestión de derechos. La ciencia y la docencia, regidas por el altruismo general, pero también por

el escepticismo organizado y la "caza de recompensas"[6] se materializan en un universo de información y documentos que también mueven intereses y abarcan un amplio rango de actividades comerciales, industriales e institucionales. Las bases de datos originadas en ese universo hacen oídos sordos a las necesidades de información de los usuarios porque cierran los ojos al contenido de las obras que representan.

3. Las causas de la ceguera

Nuestra querencia por los iconos y la comodidad de los epónimos nos impulsan a referirnos a las leyes de Newton, la enfermedad de Parkinson, el telescopio Hubble, el diccionario Bompiani o el índice de **Price**. Pero la ciencia ha sido y es, cada vez más, un producto colectivo[7] y los libros científicos o simplemente especializados también. Del casi millón y medio de libros ingresados cualquiera de estos años por las bibliotecas de universidades españolas, muchos contienen la abreviatura Ed. o sus variantes en el registro de sus menciones de responsabilidad: se producen mediante una agregación de capítulos de autores diversos coordinados por uno o más responsables. Quien busque en un catálogo la tecnología de la panificación se verá recompensado por el título *Technology of Breadmaking*. Quien desee informarse sobre el retardo y la congelación de la masa [panaria] desistirá de la consulta a ese mismo catálogo, a pesar de que las casi 30 páginas que el mismo libro contiene sobre esa tecnología específica. Es sabido que, en publicación científica, la unidad de distribución (un volumen o un fascículo) raramente coincide con la unidad de información (un capítulo o un artículo).

El reducido nivel de representación de los libros es causa de la invisibilidad de su contenido incluso en bases de datos estructuradas. Por mucho que el catálogo sea un "documento secundario que registra y describe documentos reunidos…", la consulta se realiza en una abrumadora proporción de los casos mediante el empleo del título o de los autores (88,78 y 87,59% respectivamente en el estudio de **Ortiz-Repiso** y colaboradores sobre el uso del *opac* del *Csic*[8]). En otras palabras, se realizan búsquedas de "ítem conocido" con objeto de ubicar una obra y no se busca la existencia de obras con determinado contenido (ni un 2% de los usuarios recurre a los códigos clasificatorios). Y el caso es que los capítulos de libro ocupan el segundo lugar entre los tipos de obras citadas por los trabajos de investigación.

El reducido nivel de representación de los libros es causa de la invisibilidad de su contenido incluso en bases de datos estructuradas

Siempre que esta cuestión se plantea, se argumenta que el recurso a bases de datos especializadas puede complementar las posibilidades de recuperación de la información contenida en libros. Casi nunca es así en el caso de bases de datos de literatura científica (*Food Science and Technology Abstracts, Web of Knowledge, PubMed, Scopus,…*). Casi siempre es así en el caso de los catálogos comerciales (de cualquier editorial y de muchas distribuidoras) por el simple hecho de que introducen en sus registros una descripción de las obras que quieren comercializar. Los usuarios de los grandes catálogos especializados no perciben el contenido de los libros porque los profesionales no facilitan su visualización: un caso claro de ceguera delegada, pero ceguera al fin.

4. Terapia en tres fases

Las hojas que acompañan al registro de un libro en su solicitud de *Isbn* son espantosas[9]. Sobre todo para los autores, que no parecen tener bastante con haber ideado y realizado su obra: además tienen que cumplimentarlas. Este proceso no habría de ser más doloroso si a la plétora de datos administrativos requeridos se añadiera esa maravillosa forma de resumen que se llama sinopsis. Ese pequeño texto, concebido para las solapas o las cubiertas del libro en cuestión, no sólo contiene estimulantes frases que animan a la lectura del interior; también elementos de acceso a su trapisonda, a su asunto, a aquello que

aproxima su contenido a las expectativas del posible lector. Si una simple decisión administrativa animara a incluir la sinopsis en la solicitud de registro, la base de datos y sus usuarios habrían ganado mucho. Pero eso no sería suficiente.

Imaginemos que un *Z39.50* "mejorado" complementa la norma original, que permite el entendimiento entre sistemas catalográficos y, en consecuencia, la incorporación automatizada a un catálogo de registros procedentes de otro.

Supongamos que ese complemento o ampliación permite la cómoda extracción de contenidos de obras especializadas a partir de los registros editoriales o administrativos, una especie de catalogación en publicación que abarque elementos informativos, además de puntos de acceso convencionales. Imaginemos que el *Marc 21 xml Schema*[10] se parece extraordinariamente a lo que aquí se sugiere. Eso estaría bien, pero aún no bastaría.

En tercer lugar, supongamos que quienes elaboran, revisan y actualizan normas catalográficas y los sistemas automatizados que mantienen las bases de datos bibliográficas:

– a) han reconocido la existencia del problema, y

– b) han dispuesto los medios para que los esquemas de datos catalográficos acojan esos elementos diferenciados de contenido.

Mmm, veamos… Desde hace mucho tiempo, se reconoce la existencia de las "partes componentes" como aquellas partes de una publicación que, a efectos de acceso e identificación bibliográfica, dependen de la publicación genérica que la incorpora[11]. Desde hace algo menos, existen arcanos del estilo:

07 - *Bibliographic level*
a – *Monographic component part*

o de este otro:

505 - *FORMATTED CONTENTS NOTE (R)*
Indicators
First – Display constant controller
0 – *Contents*

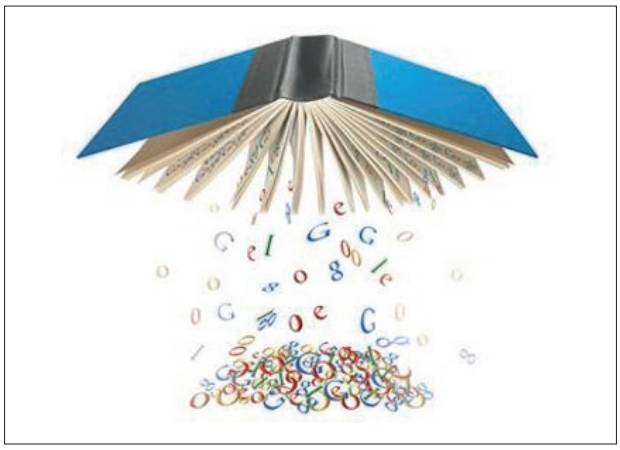
Donde se especifica que *505 – Formatted Content Note* incluye *"The titles of separate works or parts of an item or the table of contents…"*[12] ¿A que está bien?

El tratamiento completo incluiría la labor de autores y editores y no requeriría en absoluto la sobrecarga del proceso técnico en las bibliotecas. La sola voluntad de los productores pondría a disposición de los sistemas catalográficos toda una serie de elementos de acceso que, a través de la interconexión de sistemas, permitiría el acceso a las unidades de información enmascaradas hasta ahora por simples títulos genéricos.

Cuando nuevos procedimientos, patentes y algoritmos permitan el intercambio de perfiles entre buscadores y bases de datos, cabe suponer que los libros serán visibles a través de sus títulos

5. La necesidad de reacción

Mucho está cambiando sobre la recuperación de información en internet. En lo que hace poco se ha considerado una generalizada demostración de idiocia, la *Wikipedia* (a la que contribuyen alrededor de 110.000 capitidisminuidos) ofrece los primeros y quizá más acertados resultados de búsqueda en muchos sistemas. No es infrecuente que esos mismos buscadores ofrezcan resultados procedentes de otras bases de datos de absoluta solvencia. El protocolo *sitemap*[13] está posibilitando la emergencia de ésos. Además, la inventiva de los jóvenes informáticos conoce pocos límites y hace unos meses que **Jayant Madhavan**, **Zachary Ives** y **David Ko** presentaron un método para la búsqueda de contenidos que reestructura las peticiones para ajustarlas a los esquemas de datos de bases estructuradas, normalmente visualizadas como formularios[14]. **Monika Henzinger** acaba de ofrecer un breve atisbo de otros desarrollos, que revelan la reacción del departamento de investigación de *Google* al problema de la invisibilidad de registros y otros similares[15]. Cuando nuevos procedimientos, patentes y algoritmos permitan el intercambio de perfiles entre buscadores y bases de datos, cabe suponer que los libros serán visibles a través de sus títulos. Pero, ¿también lo serán sus contenidos?, ¿son iniciativas como *GoogleBooks* y su programa de bibliotecas o



los libros a texto completo de *PubMed* la única solución?

Los contenidos librarios no sólo tienen que ser visibles para los usuarios de los buscadores. Las bases de datos bibliográficas y los sistemas catalográficos tienen su propia carta de naturaleza, pero sus usuarios aún no se pueden desprender del bastón: el que un lector demande un capítulo aparentemente ilocalizable que, sin embargo, se agazapa en las páginas de uno de los libros expuestos como novedades no es un episodio infrecuente, sólo chocante (de chocar). Para que esos encontronazos queden apartados de la práctica habitual de las bibliotecas, generales o especializadas, es necesario seguir el camino que se proponía en el apartado terapéutico de este opúsculo:

– 1) la potenciación de la base de datos *Isbn* y sus sistemas de registro;

– 2) la explotación de formatos semiestructurados de intercambio, tanto desde el punto de vista de los productores como de los sistemas catalográficos; y

– 3) un grado de reconocimiento de los problemas señalados, quizá el peor de los requisitos, porque estaría asociado a una voluntad de cambio poco extendida entre profesionales.

Valdría la pena dejar de acumular garrapatas para empezar a ofrecer escarabajos de oro.

Referencias

1. **Amat, C. B.** "Rendimiento de ocho sistemas de recuperación de información en el espacio Web español". En: *El Profesional de la Información*, 2005, v. *14, n.* 5, pp. 335-346.

2. **Bergman, M. K.** "The deep web: surfacing hidden value". En: *Journal of Electronic Publishing*, 7. Recuperado de: *http://www.press.umich.edu/jep/07-01/bergman.html*

3. **Aguillo, I.** "Internet invisible o Infranet: definición, clasificación y evaluación". *Fesabid. Séptimas Jornadas Españolas de Documentación*, Bilbao, 19 a 21 de octubre de 2000.

4. *http://www.mcu.es/comun/bases/isbn/ISBN.html*

5. *http://www.mcu.es/bibliotecas/MC/CBPE/index.html*

6. **Strevens, M.** "The role of the priority rule in Science". En: *Journal of Philosophy,* 2003, v. 100, n. 2, pp. 55-79.

7. **Wuchty, S.**; **Jones, B. F.**; **Uzzi, B.** "The increasing dominance of teams in production of knowledge". En: *Science*, 2007, v. 316, n. 5827, pp. 1036-1039.

8. **Ortiz-Repiso, V.**; **Bazán, V.**; **Ponsati, A.**; **Cottereau, M.** "How researchers are using the *OPAC* of the Spanish Council for Scientific Research Library Network. En: *The Electronic Library,* 2006, v. 24, n. 2, pp. 190-211.

9. *http://www.mcu.es/libro/docs/ImpresoSolicitudISBN.doc*

10. *http://www.loc.gov/standards/marcxml/*

11. *IFLA*. Guidelines for the application of ISBDs to the description of Component Parts. Washington, Cataloging Directorate, Library of Congress, 2003. Recuperado de: *http://www.ifla.org/VII/s13/pubs/Component_Parts_final.pdf*

12. *MARC 21 Concise Format for Bibliographic Data. 2006 Concise Edition. Update 7*. Washington, Network Development and *MARC* Standards Office. Recuperado de: *http://www.loc.gov/marc/bibliographic/*

13. *http://www.sitemap.org*

14. **Halevy, A. Y.**; **Madhavan, J.**; **Ko, D. H.** Searching through content which is accessible through web-based forms. US Patent application 20060230033, 2006.

15. **Henzinger, M.** "Search technologies for the internet". En: *Science*, 2007, v. 317, n. 5837, pp. 468-471.